MA 214 Lecture 07 - 08

- Range of slopes that we can get.
- Inferential questions
  - The intercepts and the slope.
- Data descriptions
  - Example: As x change, how is y being impacted?
    - For every 1 year increase in the amount of experience a person has, the average
- Predictions
  - Use input with an statistics model, to predict the output.
    - Know x, want to know y.
- Control problems
  - Know y, want to know x.

$Y = 125$

$125 = -0.675 + 14.229x$

$x = \frac{125 + 0.675}{14.229}$

Salary = -0.7 + 14. Experience.

- Model only valid in its range of the x values.
  - Extrapolation.

- Danger of Extrapolation.

- Some useful definitions
  - Predicted values or fitted values
    - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, 2, 3, \ldots, n$
  - Residuals
    - $e_i = y_i - \hat{y}_i$, $i = 1, 2, 3, \ldots, n$.
  - Error Sum of Squares (SSE).
    - $\sum_{i=1}^{n} e_i^2$

- Residuals in "other terms"
  - $x_1, x_2, x_3, \ldots, x_n$
  - $\bar{x}$, mean (average)

Residuals in this case is:

$(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots (x_n - \bar{x})$

Estimation of Error Variance σ^2

$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{SSE}{n-2}$

$\frac{SSE}{n-2} = MSE$

$\frac{(-2)^2 + 3^2 + 2^2 + 0^2 + (-3)^2}{3}$

Mean Squared Error (MSE)

- We use the sample data to approximate the population mean.
  - Loss of the degrees of freedom

SSE is degrees of freedom is $(n-2)$

- Example
  - Fitted value: 8
  - Residual: $y - \hat{y} = 9 - 8 = 1$
  - Actual value: 9.

- If the computations are done correctly, the residuals are always added up to zero.

- $\hat{\sigma}^2 = \frac{1}{n-2}\left[\sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i\right]$

Example: Can the values of $\{-1.2, 2.1, 1.2, -1\}$

- IVu. The sum of residuals has to be added up to zero.

Inference about model parameters ★ (Hypothesis testing procedures).

- Inference about β_1 (Slope)
  - To test $H_0: \beta_1 = b$ (most of the time, $b = 0$)

H_0: There is no linear association between X and Y.

H_a: There exists a linear association between X and Y.

Monetarily. $H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

Test statistic:

$t \quad \frac{\hat{\beta}_1 - b}{S(\hat{\beta}_1)}$. where $S(\hat{\beta}_1)$ is the standard error of the estimate $\hat{\beta}_1$ and is given by:

$DF = n-2$.

$$S^2(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- So, calculation process:
  - $S^2(\hat{\beta}_1)$
  - $S(\hat{\beta}_1)$ ← we take a square root of $S^2(\hat{\beta}_1)$
  - $t = \frac{\hat{\beta}_1 - b}{S(\hat{\beta}_1)}$

Example:

Reject H_0 if $t > t_{0.025}$ or $-t_{0.025}$ with 10 df

Example: Test the hypothesis $H_0: \beta_1 = 0$ vs. $H_a: \beta \neq 0$ at 5% of significance.

"General t-test form"

$H_0: \mu = \mu_0$

then we will use: $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ Student error.

- In a sense, "the intercept" value may not have any practical meaning most of the times.

Inference about the dependent variable. (Confidence Intervals)

- Two types of inferences.
  - (Estimation of Mean Problem)
  - Inference about the mean of y, expected value of y or $E(y)$ at a given value of $x = x_0$ Average, or "everyone"
  - Prediction of an individual value of y at a given value of $x = x_p$. "Capture a target in a window..." "Particular", "Individual instance".

Confidence Intervals for Inference about model parameters.

For β_1 with confidence coefficient $(1-\alpha)$ is given by:

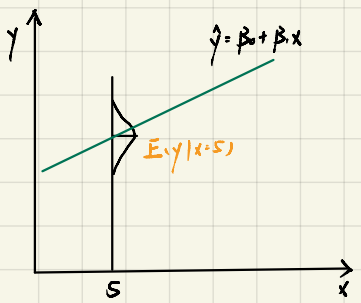
$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}} \cdot S(\hat{\beta}_1)$$

$t_{\frac{\alpha}{2}}$ is based on $(n-2)$ degrees of freedom.

$\hat{\beta}_1 = 14.229$, $S(\hat{\beta}_1) = 2.553$, $t_{0.025} = 2.228$ (tbl).

$\leftrightarrow (8.541, 19.92)$

Example:



Example: Construct a 95% confidence interval for $E(Y)$ at $x = 5$.
This is an estimate of the average annual salary for all the people.

Interval estimation of $E(Y)$ at $x = x_m$

Point Estimate of $E(Y)$: $\hat{y}_m = \hat{\beta}_0 + \hat{\beta}_1 x_m$.

Confidence Interval: $\hat{y}_m \pm t_{\frac{\alpha}{2}} \cdot s(\hat{y}_m)$.

$$s(\hat{y}_m) = \hat{\sigma}\sqrt{\quad}$$

"Mean problem"

- $E(X=5)$ and $E(X=9)$ in this case, the sample mean $\bar{x} = 5.5$
  - 5 will have narrower width for the confidence interval than the 9.
  - 9 is closer to the boundary.
    - Pursuing less useful info than 5.

"Prediction problem"

"My best guess is always the average".

The best prediction of a random predictor is always the mean.

Prediction interval of Y at $x = x_p$

Best Prediction of Y: $\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$

Standard Error of $\hat{y}_p$.

$$s(\hat{y}_p) = \hat{\sigma}\sqrt{\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

Note: $s^2(\hat{y}_p) = \hat{\sigma}^2 + s^2(y_m)$

Prediction interval of Y at $x = x_p$

$$\hat{y}_p \pm t_{\frac{\alpha}{2}} s(\hat{y}_p)$$

Regression Analysis: Part II

- Quantify how well the model fits the data.
- Judge if the fitted regression model is an adequate description of the data.

- "How far the data points are from the mean?"
  - The distance from the data points to the mean is "error", or residual.

$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$ and Recall: $SSE = \sum_{i=1}^{n} e_i^2$

this is the amount of errors with the "x information"

$SSR = SSTO - SSE$

Regression Sum of Squares.

Analysis of Variance Approach to Regression

Overall Variability in y (SSTO) → Part Explained by Regression Model (SSR)

Overall Variability in y (SSTO) → Part remains unexplained by regression (SSE)

Let's say: $SSTO = 10$

$SSE = 8$

then $SSR = 2$

The concept of "How many percent of the model can be explained?"